International Journal of Technological Exploration and Learning (IJTEL)
ISSN: 2319-2135
Volume 2 Issue 5 (October 2013)

Scheduled Theoretical Restoration for Mining Immensely Partial Data Sets

K.A.VarunKumar
Department of Computer Science and Engineering
Vel Tech Dr.RR & Dr.SR Technical University, Chennai

S.Sibi Chakkaravarthy
Department of Computer Science and Engineering
Vel Tech Dr.RR & Dr.SR Technical University, Chennai

M.Prabakaran, Ajay Kaurav
Department of Electrical & Electronics Engineering
Vel Tech Dr.RR & Dr.SR Technical University, Chennai

R.Baskar, N.Nandhakishore
Department of Electrical & Electronics Engineering
Vel Tech Dr.RR & Dr.SR Technical University, Chennai

Abstract— **Partial data sets have turn out to be just about ubiquitous in an extensive range of application fields. Mutual illustrations can be initiate in climate and image data sets, sensor data sets, and medical data sets. The partiality in these data sets may stand up from a number of issues: In some circumstances, it may merely be a replication of definite measurements not being obtainable at the time, in others, the data may be absent due to incomplete system failure, or it may merely be a consequence of users being reluctant to stipulate attributes due to confidentiality worries. When a important portion of the entries are lost in all of the attributes, it turn into very tough to perform any generous of sensible extrapolation on the unique data. For such circumstances, we present the innovative idea of theoretical restoration in which we make effective theoretical representations on which the data mining algorithms can be openly smeared. The desirability behind the idea of theoretical restoration is to practice the correlation structure of the data in directive to precise it in terms of ideas rather than the unique dimensions. As a outcome, the restoration procedure evaluates only those theoretical aspects of the data can be mined from the partial data set, rather than might faults formed by extrapolation. We reveal the efficiency of the method on a range of actual data sets.**

***Keywords*—Data Mining; Correlation; Partial Data Sets; Resortation..**

I. INTRODUCTION

In recent years, a huge number of data sets which are obtainable for data mining asks are somewhat particular. A partly specified information is one in which a firm proportion of the ideals are lost. This is because the information set for data mining troubles are usually extract from real-world situation in which also not all capacity may be offered or not all the entry may be significant to a given testimony. In other cases, where report is obtained from user straightforwardly, many users may be averse to give the entire attribute because of seclusion concern [3], [16]. In many cases, such situation upshot in records sets in which a huge profit of the ingress are lost. This is a difficulty most data mining algorithms imagine that the data set is entirely specified.

There are ranges of solution which can be used in tidy to grip this inequality for data mining especially unfinished data sets. For example, if the incompleteness occurs in a small number of rows, then such rows may be unobserved. Alternatively, when the incompleteness occurs in a small digit of column, then only these columns may be unobserved. In many cases, this summary data set may be sufficient for the reason of a data mining algorithm. None of the above technique would work for a data set which is especially imperfect because it would lead to ignoring almost of the report and attributes. General solutions to the omitted data problem include the use of assertion, arithmetical or regression-based events [4], [5], [10], [11], [19], [20], [15], [17] in order to opinion of the entries. Unfortunately, these techniques are also flat to inference errors with increasing dimensionality and incompleteness. This is because, when enormous profits of the entries are missing, each quality can be estimated to a much lower level of accuracy. Additionally, some attributes can be anticipated to a much inferior degree of reliability than others and there is no method of easy-to-read a priori which estimation are the most precise. A conversation and examples of the nature of the preconception in use straight imputation-based events may be found in [7].

We note down that any lost data machine would rely on the fact that the quality in a data set are not sovereign of one another, but that there is some extrapolative value from one characteristic to another. If the attribute in an information set are truthfully uncorrelated, then some losses in attribute entries lead to a true failure of in sequence. In such cases, absent data mechanism cannot give any approximation to the true value of a data access. Unfortunately, this is not the case in most actual data sets in which there are wide redundancies and correlation across the data demonstration.

In this paper, we discuss the novel technique of conceptual rebuilding in which we state the data in terms of the most important concepts of the link arrangement of the data. This abstract structure is gritty using techniques such as major Component Analysis [8]. These are the information in the facts along which most of the inconsistency occurs and are also referred to as the conceptual information. We note that, still a data set may contain thousands of dimensions; the numeral of concepts in it may be fairly small. For example, in text data sets, the number of size (words) is more than 100,000, but there are often only 200-400 prominent concepts [14], [9]. In this paper, we will provide verification of the claim even though predict the data along random information (such as the unique set of dimensions) is full with errors. This problem is especially true in extremely incomplete data sets in which the errors caused by successive imputation add up and result in a considerable drift from the true results. On the other hand, the mechanism along the theoretical information can be predicted quite consistently. This is because the abstract rebuilding method uses these redundancies in an effective way so as to approximation whatever abstract representations are constantly possible rather than force extrapolations on the unique set of

attributes. As the data dimensionality increases, even massively incomplete data sets can be modeled by using a small number of conceptual directions which capture the overall correlations in the data. Such a strategy is valuable since it simply tries to get anything information is really accessible in the data. We note that this results in some loss of interpretability with respect to the original dimensions; however, the aim of this paper is to be able to use available data mining algorithms in an effective and accurate way. The results in this paper are presented only for the case when the data is presented in explicit multidimensional form and are not meant for the case of latent variables.

This paper is organized as follows: The remains of this section provide a formal discussion of the offerings of this paper. In the next section, we will discuss the basic Conceptual reconstruction procedure and provide intuition on why it should work well. In Section 3, we provide the completion details. Section 4 contains the empirical results. The conclusions and summary are contained in Section 5.

A. *Contributions of this Paper*

This paper discusses a method for mining extremely incomplete data sets by exploit the connection structure of data sets. We use the connection behavior in order to create a new symbol of the data which predict only as a good deal information as can be dependably predictable from the data set. This outcome in a new full-dimensional demonstration of the data which does not have a one-to-one mapping with the unique set of attributes. However, this new representation reflects the available concept in the data correctly and can be used for many data mining algorithms, such as cluster, similarity search, or cataloging.

II. AN INTUITIVE UNDERSTANDING OF CONCEPTUAL RECONSTRUCTION

In order to make easy further discussion, we will define the profit of attribute missing from a data set as the incompleteness factor. The elevated the incompleteness factor, the trickier it is to obtain any important arrangement from the data set. The abstract reform technique is customized toward data mining particularly incomplete data sets for high-dimensional evils. As designate earlier, the attribute in high-dimensional data are often unified. This results in a normal conceptual structure of the data. For example, in a market storage bin application, a concept may consist of groups or sets of confidentially correlated items. A given client may be involved in particular kinds of items which are correlated and may vary over time. However, her conceptual performance may be much clearer at a collective level since one can classify the kinds of items that she is most interested. In such cases, even when an enormous fraction of the attributes are missing, it is possible to obtain an idea of the conceptual performance of this customer.

A more scientifically exact method for finding the collective conceptual information of a data set is Principal Component Analysis (PCA) [8]. Consider a data set with N records and dimensionality d. In the first step of the PCA technique, we generate the covariance matrix of the data set. The covariance matrix is a d * d matrix in which the (i,j),the entry is equal to the covariance among the dimensions I and j. In the second step, we generate the eigenvectors fe1 . . . edge of this covariance matrix. These are the information in the data which are such that, when the data is probable along this information, the second order connection is zero. Let us assume that the Eigen value for the eigenvector $\overline{e_i}$ is equal to λ_i. When the data is misshapen to this new Axis-system, the value is also equal to the difference of the data along the axis $\overline{e_i}$. The assets of this revolution are that most of the variance is retain in a small number of eigenvectors matching to the largest values of λ_i. We retain the k < d eigenvectors which communicate to the Largest Eigen values. An important point to appreciate is that the removal of the lesser Eigen values for extremely connected high-dimensional harms results in a new data set in which much of the clamor is removed [13] and the qualitative effectiveness of data mining algorithms such as similarity search is improved [1]. This is because these few eigenvectors communicate to the conceptual information in the data along which the no noisy aspects of the data are sealed. One of the appealing results that this paper will show is that these applicable information are also the ones along which the conceptual mechanism can be most exactly predict by using the data in the area of the relevant record. We will explain this idea with the help of an example. Throughout this paper, we will refer to a retain eigenvector as a concept in the data.

A. *On the Effects of Conceptual Reconstruction*

Let Q is a record with some missing attributes denoted by B. Let the specified attribute be denoted by A. Note that, in order to estimation the conceptual component along a given direction, we find a set of neighborhood records based on the known attributes only. These records are used in order to estimate the corresponding conceptual coordinates. Correspondingly, we define the concept of an. €; A.-neighborhood of a data point Q. Once we have established the concept of. €; A.-neighborhood, we shall define the concept of €; A; e.-predictability along the eigenvector e. intuitively, the predictability along an eigenvector e is a measure of how closely the value along the eigenvector e can be predicted using only the behavior of the neighborhood set S.Q, €, A.

Since the above fraction measures the mean to normal divergence ratio, greater amount of certainty in the correctness of the forecast is obtained when the ratio is high. We note that the value of the inevitability has been defined in this way, since we wish to make the definition scale invariant. We shall now illustrate, with the help of an example, why A; e.-predictability of eigenvector e is higher when the corresponding Eigen value is larger. In Fig. 1, we have shown a two-dimensional example for the case when a data set is drawn from a uniformly distributed rectangular distribution centered at the origin. We also assume that this rectangle is banked at an angle _ from the X-axis and the sides of this rectangle are of lengths a and b, respectively. Since the data is unvaryingly generated within the rectangle, if we were to perform PCA on the data records, we would obtain eigenvectors parallel to the sides of the rectangle. The corresponding Eigen values would be relative to a2 and b2, respectively. Without loss of generality, we may assume that a > b. Let us think that the eigenvectors in the corresponding instructions are e1 and e2, respectively. Since the variation along the eigenvector e1 is bigger, it is clear that the equivalent Eigen value is also larger. Let Q be a data point for which the X-coordinate x is shown in Fig. 1. Now, the set S(Q,ϵ{X}). Of data proceedings which is nearby to the point Q based on the coordinate X. X is in a thin strip of width 2 centered at the sector marked with a length of c in Fig. 1. In order to make an instinctive study without edge effects, we will assume that $\epsilon \rightarrow 0$. Therefore, in the diagram for Fig. 1, we have just used a erect line which is a band of width zero. Then, the standard

difference of the records in $S(Q, \epsilon, \{X\})$. Along the Y axis is given by $c/\sqrt{12} = b \cdot secant(\theta)/\sqrt{12}$ using the method for a uniform giving out along a time of length c. The equivalent components along the eigenvectors e1 and e2 are $d/\sqrt{12} = |c \cdot sine(\theta)/\sqrt{12}|$ and $e/\sqrt{12} = |c \cdot cosine(\theta)/\sqrt{12}|$, respectively. The equivalent means along the eigenvectors e1 and e2 are given by $|x \cdot sec(\theta)|$ and 0, respectively. Now, we can replacement for the mean and standard difference values in definition 2 in order to obtain the following results:

1. The $(\epsilon, \{X\}, \overline{e_1})$-predictability of the data point Q is $|x/b \cdot sine(\theta)|$.

2. The $(\epsilon, \{X\}, \overline{e_2})$-predictability of the data point Q is 0.

Thus, this example illustrates that certainty is much better in the way of the bigger eigenvector e1. Furthermore, with a outline value of inevitability along this eigenvector (which has an angle _ with the particular characteristic) improve. We will now proceed to party some of these innate results.

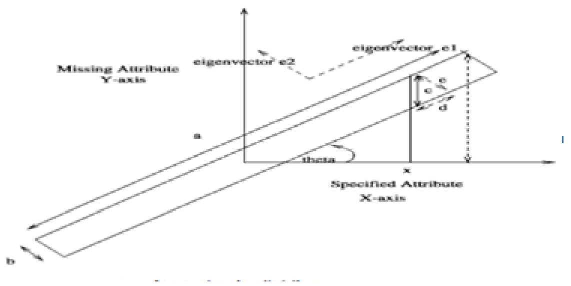

Figure 1. Pridictablity for a simple distribution

Since the above fraction measures the mean to normal divergence ratio, greater amount of certainty in the correctness of the forecast is obtained when the ratio is high. We note that the value of the inevitability has been defined in this way, since we wish to make the definition scale invariant. We shall now illustrate, with the help of an example, why A; e.-predictability of eigenvector e is higher when the corresponding Eigen value is larger. In Fig. 1, we have shown a two-dimensional example for the case when a data set is drawn from a uniformly distributed rectangular distribution centered at the origin. We also assume that this rectangle is banked at an angle _ from the X-axis and the sides of this rectangle are of lengths a and b, respectively. Since the data is unvaryingly generated within the rectangle, if we were to perform PCA on the data records, we would obtain eigenvectors parallel to the sides of the rectangle. The corresponding Eigen values would be relative to a2 and b2, respectively. Without loss of generality, we may assume that a > b. Let us think that the eigenvectors in the corresponding instructions are e1 and e2, respectively. Since the variation along the eigenvector e1 is bigger, it is clear that the equivalent Eigen value is also larger. Let Q be a data point for which the X-coordinate x is shown in Fig. 1. Now, the set S(Q,ϵ{X}). Of data proceedings which is nearby to the point Q based on the coordinate X. X is in a thin strip of width 2 centered at the sector marked with a length of c in Fig. 1. In order to make an instinctive study without edge effects, we will assume that $\epsilon \rightarrow 0$. Therefore, in the diagram for Fig. 1, we have just used a erect line which is a band of width zero. Then, the standard difference of the records in $S(Q, \epsilon, \{X\})$. Along the Y axis is given by $c/\sqrt{12} = b \cdot secant(\theta)/\sqrt{12}$ using the method for a uniform giving out along a time of length c. The equivalent components along the eigenvectors e1 and e2 are $d/\sqrt{12} = |c \cdot sine(\theta)/\sqrt{12}|$ and $e/\sqrt{12} = |c \cdot cosine(\theta)/\sqrt{12}|$, respectively. The equivalent means along the eigenvectors e1 and e2 are given by $|x \cdot sec(\theta)|$ and 0, respectively. Now, we can replacement for the mean and standard difference values in definition 2 in order to obtain the following results:

1. The $(\epsilon, \{X\}, \overline{e_1})$-predictability of the data point Q is $|x/b \cdot sine(\theta)|$.

2. The $(\epsilon, \{X\}, \overline{e_2})$-predictability of the data point Q is 0.

Thus, this example illustrates that certainty is much better in the way of the bigger eigenvector e1. Furthermore, with a outline value of inevitability along this eigenvector (which has an angle _ with the particular characteristic) improve. We will now proceed to party some of these innate results.

B. *Key Intuitions*

1) *Intuition*

The larger the value of the eigenvalue λ_i for $\overline{e_i}$, the superior the relative predictability of the conceptual component along $\overline{e_i}$.

These intuitions summarize the implications of the example discussed in the previous section. In the previous example, it was also clear that the level of correctness with which the conceptual component could be predicted along an eigenvector was dependent on the angle with which the eigenvector was bank with the axis. In order to formalize this notion, we introduce some additional notations. Let b1, . bn. correspond to the unit direction vector along a principle component (eigenvector) in a data set with n attributes. Obviously, the larger the value of bi, the more the variance of the outcrop of attribute i along the rule component i and vice versa.

2) *Intuition*

For a given vector ei, the larger the weighted ratio the greater the relative predictability of the conceptual component along ei.

$$\sqrt{\sum_{i \in A} b_i^2} / \sqrt{\sum_{i \in B} b_i^2},$$

I. DETAILS OF THE CONCEPTUAL RECONSTRUCTION TECHNIQUE

In this section, we outline the overall conceptual rebuilding procedure along with key implementation details. More specifically, two fundamental problems with the implementation need to be discussed. In order to find the conceptual directions, we first need to create the covariance matrix of the data. Since the data is massively incomplete, this matrix cannot be directly computed but only estimated. This needs to be carefully thought out in order to avoid bias in the process of formative the conceptual directions. Second, once the conceptual vectors (principal components) are found, we will work out the best methods for finding the components of records with missing data along these vectors.

Step 1: Compute Covariance Matrix M from data set D.
Step 2: Compute Eigenvectors $\{e_1 \dots e_d\}$ of Covariance matrix M with eigenvalues $\lambda_1 \geq \lambda_2 \dots \geq \lambda_d$.
Step 3: Retain the subset of eigenvectors $\{\overline{e_1} \dots \overline{e_m}\}$ with largest values of λ_i.
Step 4: For each record Q in D with specified attributes A and missing attributes B
Step 4A: For each retained eigenvector $\overline{e_i}$
Step 4A1: Let Y_A^i be the projection of known attribute set A of Q on $\overline{e_i}$
Step 4A2: Compute K records C which are closest to Q using the Euclidean distance on the attribute set A
Step 4A3: Let Y_B^i be the average projection of attribute set B of the records in C on $\overline{e_i}$
Step 4A4: Set the conceptual coordinate along $\overline{e_i}$ of data record Q to $Y_A^i + Y_B^i$

A. *The Conceptual Reconstruction Algorithm*

The overall conceptual reconstruction algorithm is illustrated Fig. 2. For the purpose of the following explanation, we will assume, without loss of generality, that the data set is centered at the origin.

The goal in Step 1 is to compute the covariance matrix M from the data. Since the records have missing data, the covariance matrix cannot be directly constructed. Therefore, we need methods for estimating this matrix. In a later section, we will discuss methods for computing this matrix M. Next, we compute the eigenvectors of the covariance matrix M. The covariance matrix for a data set is positive semi definite and can be expressed in the form $M = PNP^T$, where N is a diagonal matrix containing the Eigen values $\lambda_1 \dots \lambda_d$. The columns of P are the eigenvectors $\overline{e_1} \dots \overline{e_d}$, which form an orthogonal axis-system. We assume without loss of generality that the eigenvectors are sorted so that $\lambda_1 \geq \lambda_2 \geq \dots \lambda_d$. To find these eigenvectors, we rely on the popular Householder reduction to tridiagonal form and then apply the QL transform [8], which is the fastest known method to compute eigenvectors for symmetric matrices Once these eigenvectors have been gritty, we decide to retain only those which protect the greatest amount of variance from the data. Well-known heuristics for deciding the number of eigenvectors to be retained may be found in [8]. Let us assume that a total of m _ d eigenvectors $\overline{e_1} \dots \overline{e_m}$ are

B. *the Covariance Matrix*

At first sight, a normal method to find the covariance between a given pair of magnitude i and j in the data set is to simply use those entries which are specified for both dimensions i and j and calculate the covariance. However, this would often lead to substantial bias since the entries which are missing in the two dimensions are also often correlated with one another. Accordingly, the covariance between the specified entries is not a good representative of the overall covariance in a real data set. This is specially the case for particularly incomplete data sets in which the bias may be considerable. By using dimensions on a pair wise basis only, such methods ignore a substantial amount of information that is hidden in the correlations of either of these retained. Next, we set up a loop for each retained eigenvector ei and incompletely specified record Q in the database. We assume that the set of known attributes in Q is denoted by A, whereas the set of unidentified attributes is denoted by B. We first find the bulge of the specified attribute set A onto the eigenvector ei. We denote this projection by Y IA, whereas the outcrop for the indefinite attribute set B is denoted by Y iB. Next, the K nearby records to Q is determined using the Euclidean distance on the attribute set A. The value of K is a user-defined parameter and should typically be fixed to a small percentage of the data. For the purposes of our implementation, we set the value of K consistently to about 1 percent of the total number of records, subject to the restriction that K was at least 5. This envoy set of records is denoted by C in Fig. 2. Once the set C has been computed, we estimate the missing component Y iB of the projection of Q on ei. For each record in the set C, we compute its projection along ei using the attribute set B. The average value of these projections is then taken to be the estimate Y iB for Q. Note that it is possible that the records in C may also have missing data for the attribute set B. For such cases, only the components from the specified attributes are used in order to calculate the Y iB values for that record. The conceptual coordinate of the record Q along the vector ei is given by $Y^i = Y_A^i + Y_B^i$. Thus, the conceptual representation of the record Q is given by $(Y^1 \dots Y^m)$. . dimensions with the other dimensions for which fully specified values are available.

In order to strap up this hidden information, we use a procedure in which we assume a sharing model for the data and estimate the parameters of this model in terms of which the covariance are uttered. Specifically, we use the technique discussed in [10], which assumes a Gaussian model for the data and estimates the covariance matrix for this Gaussian model using an Expectation Maximization (EM) algorithm. Even though some inaccuracy is introduced because of this modeling assumption, it is still better than the vanilla approach of pair wise covariance estimation. To highlight some of the advantages of this approach, we conduct the following experiment.

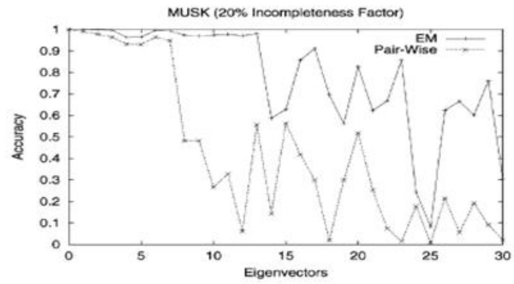

Figure 2. Comparing EM and pairwise estimation

We used the Musk data set from the UCI data set depository to create an incomplete data set in which 20 percent of the attribute values were missing. We computed the conceptual directions using both the model based approach2 and the simple pair wise covariance opinion procedure. We computed the unit direction vector (estimated vector) along each of the conceptual directions under both estimation methods and compared these direction vectors with the corresponding unit vectors constructed from the fully specified data set (actual vector). The dot product of the estimated vector and the actual vector will be in the range [0, 1], 1 indicating accident (maximum accuracy) and 0 indicating the two vectors are orthogonal (minimal accuracy). Fig. 3 describes the results of this experimentation on the first 30 eigenvectors. Clearly, the EM estimation method outperforms the pair wise estimation method. The total accuracy of the EM estimation method is also rather high. For example, for the first 13 eigenvectors (which covers more than 87 percent of the variance in the data set), the accuracy is typically above 0.94.

Once the conceptual vectors have been identified, the next step is to estimate the projection of each record Q onto each

abstract vector. In the previous section, we discussed how a set C of close records are determined using the known attributes in order to perform the rebuilding. We defined C to be the set of records in the neighborhood of Q using the attribute set A. The value for Q is estimated using the records in set C. It is possible to further refine the performance using the following observation.

The values of YB for the records in C may often show some cluster behavior. We cluster the YB values in C in order to create the sets C1 . . . Cr, where $\cup_{i=1}^{r} C_i = C$. for each set Ci, we compute the distance of its cancroids to the record Q using the known attribute set A. The cluster that is closest to Q is used to predict the value of YB. The perception behind this method is obvious.

The time complexity of the method can be obtained by summing the time required for each step of Fig. 2. The first step is the calculation of the covariance matrix, which normally (when there is no missing data) requires processing time of O (d2. N). For the missing data case, since, essentially, we use the EM procedure to estimate this matrix at each iteration until convergence is achieved, the lower bound on the total cost may be approximated as O(d2 . N. it), where it is the numeral of iterations for which the EM algorithm is run. For a more exact analysis of the intricacy of the EM algorithm and associated guarantee of convergence (to a local maximum of the log-likelihood), we refer the reader elsewhere [18], [12]. The process of Step 2 is simply the generation of the eigenvectors which requires a time of O (d3). However, since only m of these eigenvectors needs to be retained, the actual time required for the combination of Steps 2 and 3 is O$(d^2 \cdot m)$. Finally, Step 4 requires m dot product calculations for each record and requires a total time of O (N.d.m).

II. EMPIRICAL EVALUATIONS

In order to perform the testing, we used several completely specified data sets (Musk (1 & 2), BUPA, Wine, and Letter-Recognition) in the UCI3 machine wisdom storehouse. The Musk 1 data set has 475 instances and 166 dimensions. The Musk 2 data set has 6,595 instances and 166 dimensions. The Letter-Recognition data set has 16 dimensions and 20,000 instances. The BUPA data set has 6 dimensions and 345 instances. The incomplete records were generated by arbitrarily removing some of the entries from the records. We introduce a notion of incompleteness in these data sets by arbitrarily eliminating values in records of the data set. One of the advantages of this method is that, since we already know the unique data set, we can compare the effectiveness of the reconstruct data set with the actual data set to validate our approach. We use several evaluation metrics in order to test the efficiency of the reconstruction approach. These metrics are designed in various ways to test the sturdiness of the reconstructed method in preserve the inherent information from the original records.

A. Direct Error Metric

Let $Y^i_{estimated}(Q)$. be the estimated value of the conceptual component for the eigenvector i using the modernization method. Let $Y^i_{actual}(Q)$. Be the true value of the projection of the record Q on to eigenvector i, if we had an oracle which knew the true projection onto eigenvector i using the original data set. Obviously, the closer $Y^i_{actual}(Q)$ is to $Y^i_{estimated}(Q)$, the better the quality of the reconstruction. We define the relative error5 along the eigenvector i as follows:

$$Error_i = \frac{\sum_{\forall Q in D} |Y^i_{estimated}(Q) - Y^i_{actual}(Q)|}{\sum_{\forall Q in D} |Y^i_{actual}(Q)|}.$$

Clearly, inferior values of the error metric are more enviable. In many cases, even when the absolute error in estimation is somewhat high, experiential data suggests that the correlation between estimated and actual values continue to be quite high. This indicates that, even though the estimated conceptual representation is not the same as the true representation, the estimated and actual components are correlated so highly that the direct application of many data mining algorithms on the reconstructed data set is likely to continue to be effective. To this end, we computed the covariance and correlation of these actual and estimated projections for each eigenvector over different values of Q in the database. A validation of our conceptual

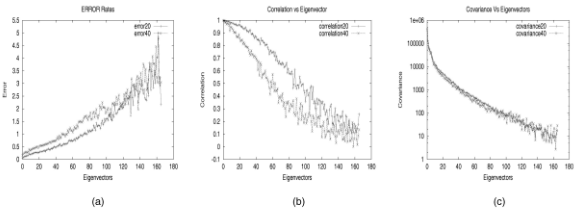

Figure 3. . (a) Error, (b) correlation (estimated, actual), and (c) covariance (estimated, actual) as a function of eigenvectors for the Musk(1) data set at 20 percent and 40 percent missing

.Rebuilding procedure would be if the correlations between the actual and estimated projections are high. Also, if the scale of the covariance between the estimated and actual mechanism along the principal eigenvectors were high, it would provide further validation of our intuitions that the principle eigenvectors provide the instructions of the data which have the maximum predictability.

B. Indirect Error Metric

Since the thrust of this paper is to compute conceptual representations for indirect use on data mining algorithms rather than actual attribute reconstruction, it is also useful to evaluate the methods with the use of an indirect error metric. In this metric, we build and compare the performance of a data mining algorithm on the reconstruct data set. To this effect, we use classifier trees generated from the original data set and compare it with the performance of the classifier trees generated from the reconstructed data set. Let CAo be the classification accuracy with the original data set and CAr is the classification accuracy with the reconstructed data set. This metric, also referred to as Classification Accuracy Metric (CAM), measures the ratio between the above two classification accuracies. More formally:

$$CAM = \frac{CA_r}{CA_o}.$$

Thus, the indirect metric measures how close to the original data set the reconstructed data set is in terms of classification accuracy.

1) Evaluations with Direct Error Metric

The results for the Musk (1) data set are shown in Fig. 4. In all cases, we plot the results as a meaning of the eigenvectors ordered by their eigenvalues where eigenvector 0 corresponded to the one with the major eigenvalue.

F4a offers some experiential evidence for Intuition 1. Clearly, the inevitability is better on eigenvectors with a superior variance. In this data set, we note that the error quickly increases for the eigenvectors with a small variance. For eigenvectors 145-165, the relative error is larger than 3. This is because these are the sound directions in the data along which there are no logical correlations among the different dimensions. For the same reason, these eigenvectors are not really relevant, even in entirely specified data sets, and are ignored from the data representation in dimensionality reduction techniques. The removal of such directions is often desirable, even in fully specified data sets, since it leads to the prune of noise effects from the data [13].

To further authenticate our approach, we calculated the covariance and correlations between the actual and estimated components along the different eigenvectors. The results are illustrated in Figs. 4b and 4c. For this data set, the leading eigenvectors show a very strong correlation and high covariance between the estimated and actual projection. The correlation value for the largest 20 eigenvectors is greater than 0:95. For the first five eigenvectors, there is about an 8 percent drop in the average error, while the correlation continues to be really significant (around 0.99).

As expected, the average errors are higher for 40 percent incompleteness factor when compared to 20 percent incompleteness factor. However, the general tendency of variation in error rate with the magnitude of the variance along a principal component is also retained in this. The correlations between the true and estimated values continue to be quite high. These results are encouraging and serve to authenticate our key intuitions, particularly given the high level of incompleteness of this data set.

Similar trends were experiential for the Musk (2), BUPA, and Wine data sets. The results are illustrated in Figs. 5, 6, and 7, respectively. Once again, for these data sets, we observed the following trends: The eigenvectors with the largest variance had the lowest estimation error, there was a very high correlation and covariance between the estimated and actual values along the eigenvectors with high variance and increasing the level of missingness from 20 to 40 percent resulted in vaguely poorer estimation quality (as determined by the direct metrics). The results for the Letter Recognition data set were slightly different and are illustrated in Fig. 8. While the observed correlations between the actual and estimated projections were sensibly high for the eigenvectors with high variance, the observed covariance were decidedly on the lower side. Furthermore, the correlations were also not quite as high as the other data sets. This is reflective of the fact that this is a data set in which the cross-attribute redundancy in data representation, i.e., the correlation structure of this data set, is weak. Such a data set is a very difficult case for the conceptual reconstruction approach or any other missing data mechanism. This is because any removal of quality values in such a data set would lead to true loss of information, which cannot be remunerated for by the inter attribute correlation redundancy. As we shall see, our experiment with the indirect metric bears this fact out.

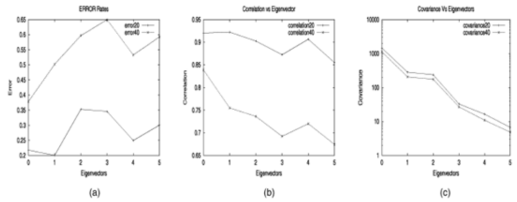

Figure 4. (a)Error,(b)correlation(estimated,actual)and(c)covariance(estimated,actual)as a function of eigenvectors for the BUPA data set at 20

However, in general, our observation across a wide range of data sets was that the correlation between the actual commechanism tends to be quite high. This strength of the correlation metric indicates that, for a particular eigenvector, the error is Usually created by either a constant underestimation or a consistent overestimation of the conceptual component. This constancy is quite significant since it implies that a simple linear translation of the origin along the eigenvector could reduce the error rate further. Of course, the direction of translation is not known a priori. However, for typical data mining tasks such as clustering and similarity search, where the relative position of the data records with respect to one another is more relevant, it is not necessary to perform this translation. In such cases, the reconstructed data set would continue to be highly reliable.

2) *Results with Indirect Metric*

Since the purpose of the conceptual rebuilding method is to provide a new representation of the data on which data mining algorithms can be straight applied, it is useful to test the effects of using the procedure on one such algorithm. To this effect, we use a decision tree classifier [19], which we apply both to the original (complete) representation and the conceptual representation of the missing data. In Table 1, we have illustrated6 the accuracy of the classifier on a conceptual representation of the data, when the percentage of incomplete entries varies from 20 to 40 percent, respectively (CAM.RC. columns). We have also illustrated the correctness on the original representation in the same table (CAO column). In addition, we also compared the reconstruction approach with a move toward that fills missing values using mean attribution (CAM.IM. columns).

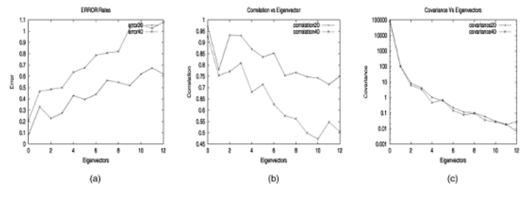

Figure 5. .(a)Error,(b)correlation(estimated,actual),and(c)covariance(estimated,actual) as a function

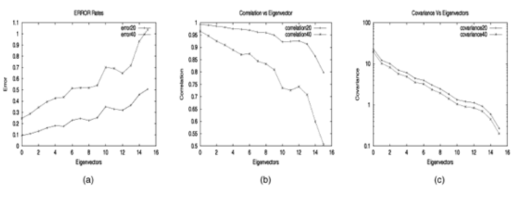

Figure 6. a)Error,(b)correlation(estimated,actual),and(c)covariance(estimated,actual)as a function of eigenvectors for the Letter-Recognition

For all the data sets and at special levels of missingness, our approach is clearly better to the approach based on mean attribution. The only exception to the above is the Wine data

set, where, at 20 percent missingness, the two schemes are equal. In fact, in some cases, the development in accuracy is nearly 10 percent. This improvement is more apparent in data sets where the correlation structure is weaker (Letter-Recognition, Bupa) than in data sets where the correlation structure is stronger (Musk, Wine data sets). One possible reason for this is that, although mean imputation often results in incorrect estimations, the stronger correlation structure in the Musk data sets enables C4.5 to ignore the incorrectly estimated attribute values, thereby ensuring that the classification performance is relatively unaffected. Note also that the improvement of our reconstruction approach over mean imputation is more noticeable as we move from 20 percent missingness to 40 percent missingness. This is true of all the data sets, including the Wine data set.

For the case of the BUPA, Musk (1), and Musk (2) data sets, the C4.5 classifier built on the reconstructed data set (our approach) was at least 92 percent as correct as the original data set, even with 40 percent incompleteness. In most cases, the accuracy was radically higher. This is evidence of the heftiness of the technique and its applicability as a procedure to transform the data without losing the inherent information available in it.

Out of the five data sets experienced, only the letter recognition data set did not show as effective a classification performance as the other three data sets. This difference is especially perceptible at the 40 percent incompleteness factor. There are three particular characters of this data set and the classification algorithm which contribute to this. The first reason is because the correlation structure of the data set was not strong enough to account for the loss of information created by the missing attributes. Although our approach outperforms mean imputation, the weak correlation structure of this data set tends to amplify the errors of the reconstruction approach. We note that any missing data mechanism needs to depend upon inter attribute redundancy and such behavior shows that this data set is not as suitable for missing data mechanisms as the other data sets. Second, on presentation the decision trees that were constructed, we noticed that, for this particular data set, the classifier happened to pick the eigenvectors with lower variance first, while selecting the splitting attributes. These lesser eigenvectors also are the ones where our estimation procedure results in larger errors. This problem may not, however, occur in a classifier in which the higher eigenvectors are picked first (as in PCA-based classifiers). Finally, in this particular data set, several of the classes are intrinsically similar to one another and are distinguished from one another by only small variations in their feature values. Therefore, removal of data values has a severe effect on the maintenance of the distinctive characteristics among different classes. This tends to increase the misclassification rate.

TABLE I. EVALUTION OF INDIRECT METRIC

Dataset	CA_o	$CAM_{20\%}(RC)$	$CAM_{20\%}(MI)$	$CAM_{40\%}(RC)$	$CAM_{40\%}(MI)$
BUPA	62.4	0.963	0.89	0.927	0.875
Musk (1)	76.2	0.943	0.937	0.92	0.89
Musk (2)	95.0	0.96	0.95	0.945	0.917
Letter Recognition	84.9	0.825	0.749	0.62	0.54
Wine	91.0	0.98	0.98	0.927	0.88

We note that, even though the applicability of the general conceptual reconstruction technique applies across the entire range of generic data mining problems, it is possible to further improve the method for picky problems. This can be done by picking or designing the method used to solve that problem more carefully. For example, we are evaluating strategy by which the overall classification performance in such reconstructed data sets can be improved. As mentioned earlier, one strategy under active consideration is to use class-dependent PCA-based classifiers. This has two advantages: First, since these are PCA-based, our reconstruction approach naturally fits into the overall model. Second,class-dependent approaches are typically better discriminators in data sets with a large number of classes and will improve the overall classification accuracy in such cases. An interesting line of future study would be to develop conceptual reconstruction approaches which are specially tailored to different data mining algorithms.

III. CONCLUSIONS AND DIRECTIONS FOR FUTURE WORK

In this paper, we present the novel idea of theoretical transformation for mining particularly incomplete data sets. The key incentive behind conceptual reconstruction is that, by choosing by prediction the data along the conceptual directions, we use only that level of knowledge that can be reliably predicted from the incomplete data. This is lither than the restrictive approach of Predicting along the original attribute directions. We show the effectiveness of the technique on a wide variety of real data sets. Our results indicate that, even though it may not be possible to reconstruct the original data set for a random feature or vector, the conceptual directions are Very agreeable to reconstruction. Therefore, it is possible to reliably apply data mining algorithms on the conceptual representation of the reconstructed data sets.

In terms of future work, one interesting line is to expand the proposed ideas to work with categorical attributes. Recall that the current approach works well only on continuous attributes since it relies on PCA. Another motivating avenue of future research could involve investigatingrefinement to the estimation procedure that can improve the efficiency (using sampling) and accuracy (perhaps by evaluating and using the refinements suggested in Section 3.1) of the conceptual reconstruction procedure.

REFERENCES

[1] C.C. Aggarwal, “On the Effects of Dimensionality Reduction on High Dimensional Similarity Search,” Proc. ACM Symp. Principles of Database Systems Conf., 2001

[2] C.C. Aggarwal and S. Parthasarathy, “Mining Massively Incomplete Data Sets by Conceptual Reconstruction,” Proc. ACM Knowledge Discovery and Data Mining Conf., 2001.

[3] R. Agrawal and R. Srikant, “Privacy Preserving Data Mining,” Proc. ACM SIGMOD, 2000

[4] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, Classification and Regression Trees. New York: Chapman & Hall, 1984.

[5]] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” J. Royal Statistical Soc. Series, vol. 39, pp. 1-38, 1977

[6] A.W. Drake, Fundamentals of Applied Probability Theory. McGraw-Hill, 1967

[7] Z. Ghahramani and M.I. Jordan, “Learning from Incomplete Data,” Dept. of Brain and Cognitive Sciences, Paper No. 108, Massachusetts Institute of Technology, 1994.

[8] I.T. Jolliffe, Principal Component Analysis. New York: Springer-Verlag, 1986.

[9] J. Kleinberg and A. Tomkins, "Applications of Linear Algebra to Information Retrieval and Hypertext Analysis," Proc. ACM Symp. Principles of Database Systems Conf., Tutorial Survey, 1999.

[10] R. Little and D. Rubin, "Statistical Analysis with Missing Data Values," Wiley Series in Probability and Statistics, 1987.

[11] R.J.A. Little and M.D. Schluchter, "Maximum Likelihood Estimate for Mixed Continuous and Categorical Data with Missing Values," Biometrika, vol. 72, pp. 497-512, 1985.

[12]] S.Parthsarthy and C.C. Aggarwal, "On the Use of Conceptual Reconstruction for Mining Massively Incomplete Data Sets, "IEEE Trans. Knowledge and Data Eng., pp. 1512-1521,2003.

[13] Eclipse Home Page : http://www.eclipse.org/

[14] http://weka.sourceforge.net/wiki/index.php/Writing_your_own_Filter

[15] wekaWikilin :http://weka.sourceforge.net/wiki/index.php/Main_Page

[16]] Ian H. Witten and Eibe Frank , "Data Mining: Practical Machine Learning Tools and Techniques" Second Edition, Morgan Kaufmann Publishers. ISBN: 81-312-0050-7.

[17] http://weka.sourceforge.net/wiki/index.php/CVS

[18] http://weka.sourceforge.net/wiki/index.php/Eclipse_3.0.x weka.filters.SimpleBatchFilter